

# Analytic queueing model for ambulance services

Pedro A. Pury

Facultad de Matemática, Astronomía, Física y Computación  
 Universidad Nacional de Córdoba  
 Ciudad Universitaria, X5000HUA Córdoba, Argentina  
 pury@famaf.unc.edu.ar

## Abstract

We present predictive tools to calculate the number of ambulances needed according to demand of entrance calls and time of service. Our analysis discriminates between emergency and non-urgent calls. First, we consider the nonstationary regime where we apply previous results of first-passage time of one dimensional random walks. Then, we reconsider the stationary regime with a detailed discussion of the conditional probabilities and we discuss the key performance indicators.

## 1 Introduction

Emergency Medical Services (EMS) involve several operational decisions concerning optimizing the station location deployment for ambulances and the selection of the number of ambulances available in the fleet at different moments of the day or on different days in the week. Quantitative and predictive tools to assist the ambulance management are becoming increasingly important in the solution of economic and medical aspects of the problem [1]. Among these tools, modeling health care problems with queueing theory has gained in significance in the last sixty years [2].

Queueing theory has been used in the study of spatial and temporal distributions of demand of EMS in order to simulate the behavior of the system. Particularly, the ambulance location problem has been subject of considerable attention (See Refs. [3, 4, 5] and references therein) and has been applied to support decision making [6]. However, the problem of the number of ambulances required for service has received little mathematical attention beyond the prediction by the Little's Law [7, 8]. The objective of this work is to derive useful quantities to predict the number of ambulances needed in real-time operation of EMS. This is an analytical work but with a didactic approach. We address here both modeling with clear mathematical derivations and their significance to operations research.

Our model is basically a call center [9] and can be described by exponential interarrival and service times, and  $s = M$  servers:  $M/M/s/GD/\infty/\infty$  [10]. Thus, our queueing model is based on two average times:  $T_C$ , the mean difference between call arrival times and  $T_S$ , the mean service time of an individual ambulance. The service time involves the complete time lapse between the dispatch of the ambulance from the base and its release for subsequent utilization. Thus, the service time sums up all the transport times of the ambulance plus the specific time for the medical attention at the scene. We are interested in the behavior of the system over a finite time interval, but long compared with  $T_C$  and  $T_S$  [11]. We focus on the nonstationary regime, where we apply the concept of the mean first-passage time (MFPT) [12], as well as in the stationary regime, where we are interested in the key performance quantities.

The paper is organized in three main sections. In Section 2 we describe the random walk model and the problem assumptions are emphasized. Section 3 focuses on the calculation of the time to the next critical condition, that is when all ambulances are busy, whereas in Section 4 we

discuss standard results of queueing theory with emphasis in the relevant performance indicators for the queue of clients and level of service of a home medical care service. Here, we also provide an alternative proof of the Little's Law. The complementary mathematical details of Sections 3 and 4 are relegated to corresponding subsections in the appendix to enhance readability. Finally, in Section 5 we briefly summarize our main results.

## 2 The random walk model

Formally, we represent the system by the total number  $M$  of ambulances in the service and by the number  $n$  of received calls not served yet.  $n$  is the state of occupation of the system. When  $n = 0$ , all the ambulances are in the base and there are not calls in queue. For  $0 < n < M$ , there are not waiting calls and  $n$  ambulances are in course of action. That is, in transit from base to the call location, attending at patient location, or in transit to the hospital. In an equivalent way, we can say that there are  $n$  patients simultaneously being served. When  $n = M$ , the system is at the critical state. Even though there are not calls in waiting, all the servers have been assigned to calls and, in consequence, there are not any ambulance available to serve an eventual next incoming call. For  $n > M$ , the system is saturated. All ambulances are occupied and there are  $(n - M)$  calls waiting to be served in the queue.

At any time, the system can change its state between its nearest neighbors. Thus, the transition probability per unit time from the state  $n$  to  $(n + 1)$  is denoted by  $\omega_n^+$ , whereas the transition probability toward the lower occupation state is given by  $\omega_n^-$ . In Figure 1, we sketch the possible transitions for the system.

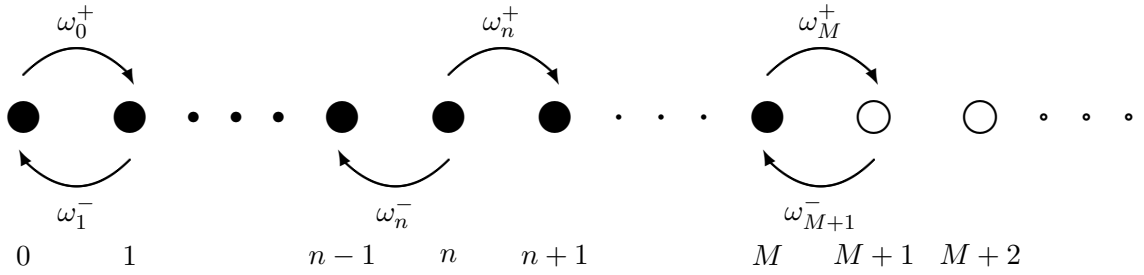


Figure 1: Scheme of transition probabilities between system's states in the upper side and the state of occupation of the system in the bottom line, for a system with  $M$  servers.

The dynamics of the probability  $P_n(t)$  of finding the system at time  $t$  in the state  $n$  is ruled by the master equations of a random walk between nearest-neighbor sites with a reflecting boundary at the origin, that is a birth-death process [10],

$$\begin{aligned} \frac{dP_0(t)}{dt} &= \omega_1^- P_1(t) - \omega_0^+ P_0(t), \\ \frac{dP_n(t)}{dt} &= \omega_{n+1}^- P_{n+1}(t) + \omega_{n-1}^+ P_{n-1}(t) - (\omega_n^+ + \omega_n^-) P_n(t) \quad (n \geq 1). \end{aligned} \tag{1}$$

We restrict our analysis to the case in which the times between calls and the service times are *independent exponential* random variables. The assumption that call arrival volume per unit of time is Poisson distributed is a standard choice in industry [13]. Thus, the time between calls is an exponential variable with rate or mean number of calls per unit time  $\lambda = 1/T_C$ . On the other hand, the service time rate or mean number of attention per unit time and per ambulance results

$\mu = 1/T_S$ . Therefore, the transition probabilities are defined by [11]

$$\begin{aligned}\omega_n^+ &= \lambda, \forall n, \\ \omega_n^- &= \begin{cases} n\mu & (n \leq M), \\ M\mu & (n \geq M). \end{cases}\end{aligned}\quad (2)$$

In this manner, only  $\omega_n^-$  depends on the state of the system. The average times  $T_C$  and  $T_S$  are all the experimental information needed to characterize the problem.

### 3 The critical emergency problem

The ambulance industry has the goal of provide care within 8 minutes for heart attack and cardiac arrest [14] and major trauma [15] (for critical considerations see Ref. [16]). Usually, the time lapse between picking the call up and the arrival of the ambulance at the scene is mainly consumed in transport. Thus, to achieve this goal is necessary to respond the calls immediately, without putting any EMS call in queue. Therefore, the most important aspect of emergency medical management is avoiding the saturation of the system. The prediction of the critical condition ( $n = M$ ) is the particular interest for the quality of the medical service as much as the economic management of the service, given that the critical condition strongly depends on the number  $M$  of ambulances simultaneously in service.

The mean time to the next critical condition is function of the initial state of the system, and can be calculated as the MFPT of a one dimensional random walk to site  $M$  with a reflecting boundary at the opposite extreme (see Figure 1). For the critical problem, the initial condition  $n$  is restricted to the values in the interval  $[0, \dots, M]$ . MFPT is a dynamical variable that can not be computed in the steady state. Using known results in the literature [17], the mean time to the critical condition as function of the initial state of the system results

$$\begin{aligned}T(0) &= T_C \left( M + 1 + \sum_{k=0}^{M-1} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^M i! \gamma^i \right), \\ T(1) &= T(0) - T_C, \\ T(n) &= T(0) - T_C \left( n + \sum_{k=0}^{n-2} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i! \gamma^i \right) \quad (2 \leq n \leq M),\end{aligned}\quad (3)$$

where the parameter  $\gamma = \mu/\lambda = T_C/T_S$ . The mathematical details are given in Appendix 6.1.

Particularly,  $T(M)$  is the mean time between two critical conditions of the system, but without reaching saturation. Furthermore, as we want a quantity independent of the initial state, we average over  $n = 0, \dots, M$ . For this purpose, we define

$$\langle T \rangle = \frac{1}{M+1} \sum_{n=0}^M T(n). \quad (4)$$

In this way, knowing the values of  $T_C$  and  $T_S$ , the expressions of Eq. (3) can be numerically evaluated in a very direct way. In Figure 2 we show the plots of  $\langle T \rangle$ , according to Eqs. (3) and (4), as function of the mean time between calls  $T_C$ . We have sketched a characteristic situation where the mean service time is 50 min and we considered the number of ambulances  $M = 5, \dots, 9$ . The curves clearly show the non-linear behavior of  $\langle T \rangle$ .

Emergency medical care is an activity brought completely under protocol and in consequence the time of paramedic duties on scene have little variance. However, in a modern metropolis, transport time could consume an important part of the total service time and it have an specific hourly pattern. But, in practical situations,  $T_S$  has small variations over the day in comparison

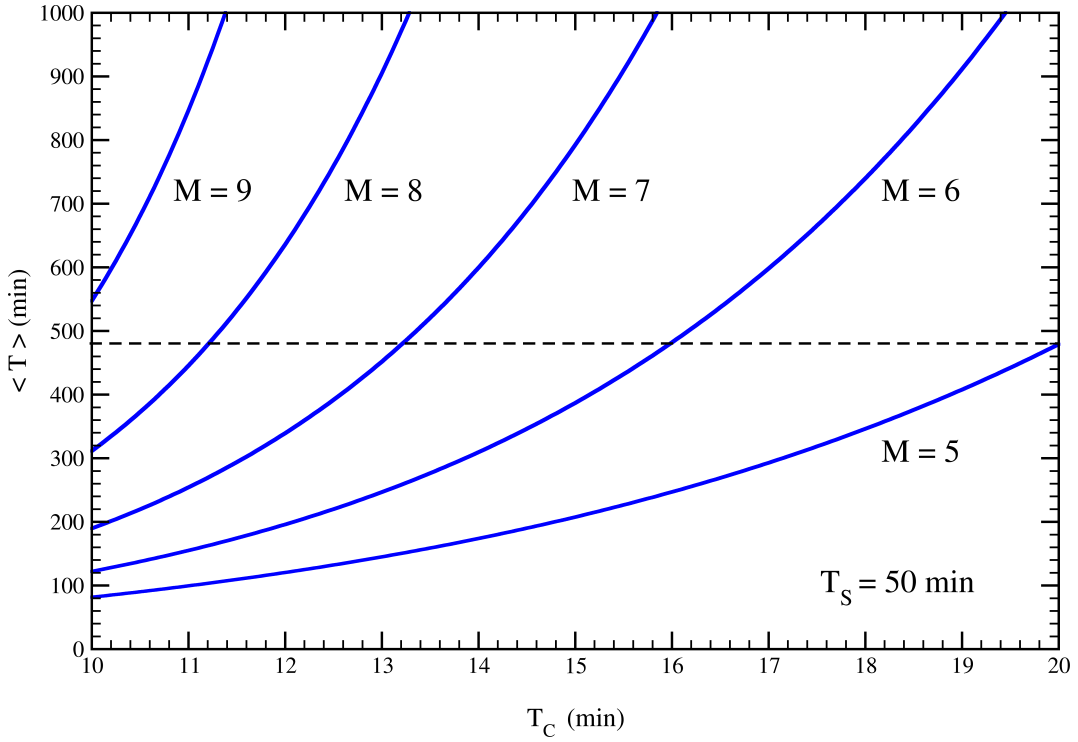


Figure 2: Mean time to critical condition as function of the mean time between incoming calls.

with  $T_C$ . The fluctuations of the rate of entrance calls,  $\lambda$ , have a defined pattern with marked differences during the day. Usually, the mean number of EMS calls per hour remains high during a time span of approximately eight hours between noon and 8:00 p.m. [13]. This distinct shape is similar for all days of the week.

Considering that  $T_C$  is measured in the time interval of highest number of calls, the mean time to critical condition must be longer than this time span to ensure the service's quality and avoiding saturation. For example, in Figure 2 we see that the system does not reach the critical condition within the interval of eight hours of high  $\lambda$  (horizontal dashed line) for  $T_C > 16$  min with six ambulances, and for  $T_C > 13.2$  min with seven ambulances.

In a similar way, to optimize the number of ambulances simultaneously in service, our analysis can be reproduced for any time interval of the day, for different days of the week, and at different locations of ambulance bases. Usually, the call rate  $\lambda$ , measured in a given interval during the day, is a linear function of the numbers of affiliates of the EMS but with seasonal fluctuations. This estimation gives us a reference for the average value of  $T_C$  in the long term. In the short run, using the forecasting of the volume of call arrival [18, 13], our analysis allows us to estimate the diary and hourly demand of ambulances. Thus, we can easily design a simple and precise predictive tool for the fleet of ambulances required for a EMS by combining the use of estimation or forecasting of  $T_C$  with our analysis of Figure 2, tailored for each particular case.

## 4 The non-urgent call problem

Most of the ambulance service providers, besides dealing with emergency calls, also bring medical home assistance for non-urgent calls. Usually, if a call is evaluated as non-urgent, the dispatcher derives it to another system to alleviate the use of emergency ambulances. In this case, contrary to the emergency management, non-urgent calls are allowed occasionally to be driven under saturation

( $n > M$ ). Thus, the length of the queue and the waiting time of the patients are the quantities of interest for the ongoing process. However, the system is driven in steady state only under a special condition.

#### 4.1 Steady state

First, we define the dimensionless control parameter

$$\rho = \frac{\lambda}{M\mu}, \quad (5)$$

which is also called *traffic intensity* in queueing theory [11]. The system has stationary state if and only if  $\rho < 1$ . Under this condition, the limit  $\pi_n = \lim_{t \rightarrow \infty} P_n(t)$  exists and according to the analysis of Appendix 6.2 is given by

$$\pi_n = \frac{1}{S} \begin{cases} \frac{M^n}{n!} \rho^n & (0 \leq n \leq M), \\ \frac{M^M}{M!} \rho^n & (n \geq M), \end{cases} \quad (6)$$

where

$$S = \sum_{n=0}^{M-1} \frac{(M\rho)^n}{n!} + \frac{(M\rho)^M}{M!(1-\rho)}. \quad (7)$$

In Figure 3 we show plots for the probability distribution, according to Eqs. (6) and (7). Left panel corresponds to a fleet of five ambulances, whereas the right panel is for seven servers.

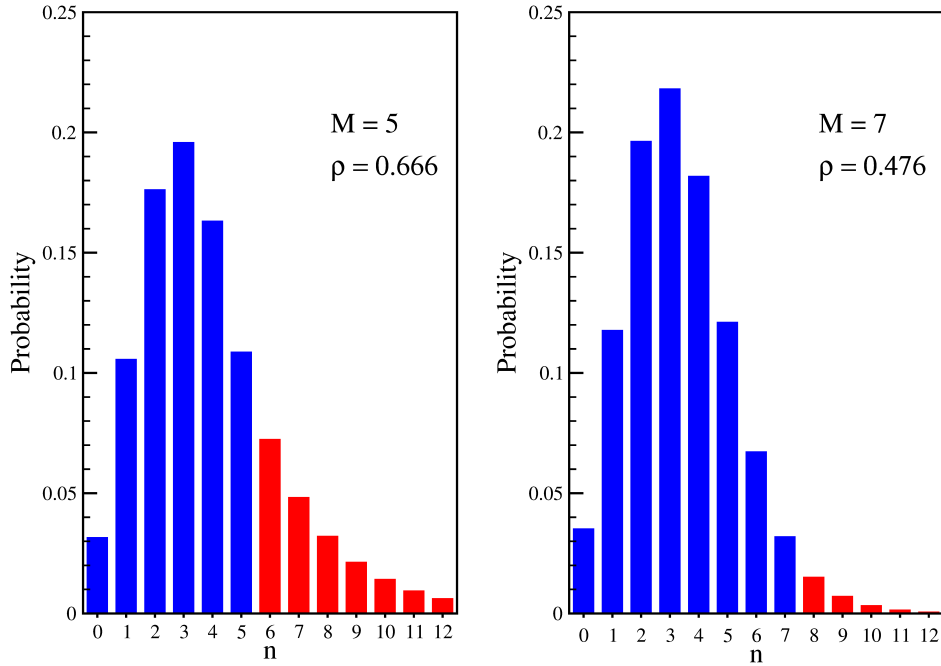


Figure 3: Probability distribution in the steady state for  $T_S = 50$  min and  $T_C = 15$  min.

When the condition  $\rho < 1$  is not fulfilled, the probabilities  $\pi_n$  are not defined. This situation corresponds to the collapse of the system when it is unable to deal with calls as fast as they arrive and the queue of waiting calls will develop without limit.

## 4.2 Queue length

If a call arrives when all servers are busy, it will be put in queue. The probability of full occupation of servers results

$$P(\text{occup}) = P(n \geq M) = \sum_{n=M}^{\infty} \pi_n = \frac{1}{S} \frac{M^M}{M!} \sum_{n=M}^{\infty} \rho^n, \quad (8)$$

and under the condition  $\rho < 1$ , using Eq. (38), we obtain

$$P(\text{occup}) = \frac{(M\rho)^M}{M! (1-\rho) S}. \quad (9)$$

The red tails in Figure 3 are the probabilities of full occupation for  $M = 5$  (left) and  $M = 7$  (right) ambulances, respectively.

To calculate the length of the genuine queue formed when the  $M$  servers are busy, we need the conditional probability that the system is at state  $n$ , given that all servers are occupied. From Eq. (6) (for  $n \geq M$ ) and (9) we obtain

$$P(n|\text{occup}) = \rho^{n-M} (1-\rho) \quad (n \geq M). \quad (10)$$

Setting  $n = k + M$ ,  $k = 0, 1, \dots$  results that the conditional probability of having  $k$  calls in the waiting row under full occupation is

$$P(k|\text{occup}) = \rho^k (1-\rho) \quad (k \geq 0). \quad (11)$$

This is the geometric distribution with parameter  $1-\rho$ . Therefore, the average length of the genuine queue,  $\langle L \rangle = \sum_{k=0}^{\infty} k P(k|\text{occup})$ , and its standard deviation,  $\sigma_L$ , result

$$\langle L \rangle = \frac{\rho}{1-\rho}, \quad \sigma_L = \frac{\sqrt{\rho}}{(1-\rho)}. \quad (12)$$

Both quantities evidently diverge in the limit  $\rho \rightarrow 1$ . To visualize the effect of the number of ambulances, we plot in Figure 4 the expressions of Eq. (12) for  $M = 4$  ( $\rho = 0.833$ ),  $\dots$ ,  $10$  ( $\rho = 0.333$ ). In this example, for  $M < 4$  results  $\rho > 1$ . We can see that to increase the fleet in more than six ambulance has not any practical consequence in the queue length.

## 4.3 Quality of service

Now, we consider the arrival of a new call when the state of the system is  $n = M + k$ , i.e., all servers are busy and there are  $k \geq 0$  calls in waiting. Under the queue discipline *first-come, first-served*, the time that the new patient will have to wait until an ambulance will be dispatched to his or her location, will be the sum of the waiting times of  $N = k + 1$  patients: The first  $k$  in the row plus the time of any of the patients in service at the arrival time.

The probability distribution of the sum of  $N$  exponential independent random variables with parameter  $\alpha$  is a Gamma distribution with parameters  $(N, \alpha)$  [19],

$$f(t; N, \alpha) = \alpha e^{-\alpha t} \frac{(\alpha t)^{N-1}}{(N-1)!}. \quad (13)$$

Thus, the conditional density of probability that a patient will be waiting a time  $t$  given  $k$  patients ahead in the row is given by Eq. (13) with  $\alpha = M\mu$  and  $N = k + 1$ . We also know that the probability of having  $k$  patients in the row is given by Eq. (11). Therefore, the probability density function of the waiting time in the row results

$$g(t) = \sum_{k=0}^{\infty} (1-\rho) \rho^k f(t; (k+1), M\mu). \quad (14)$$

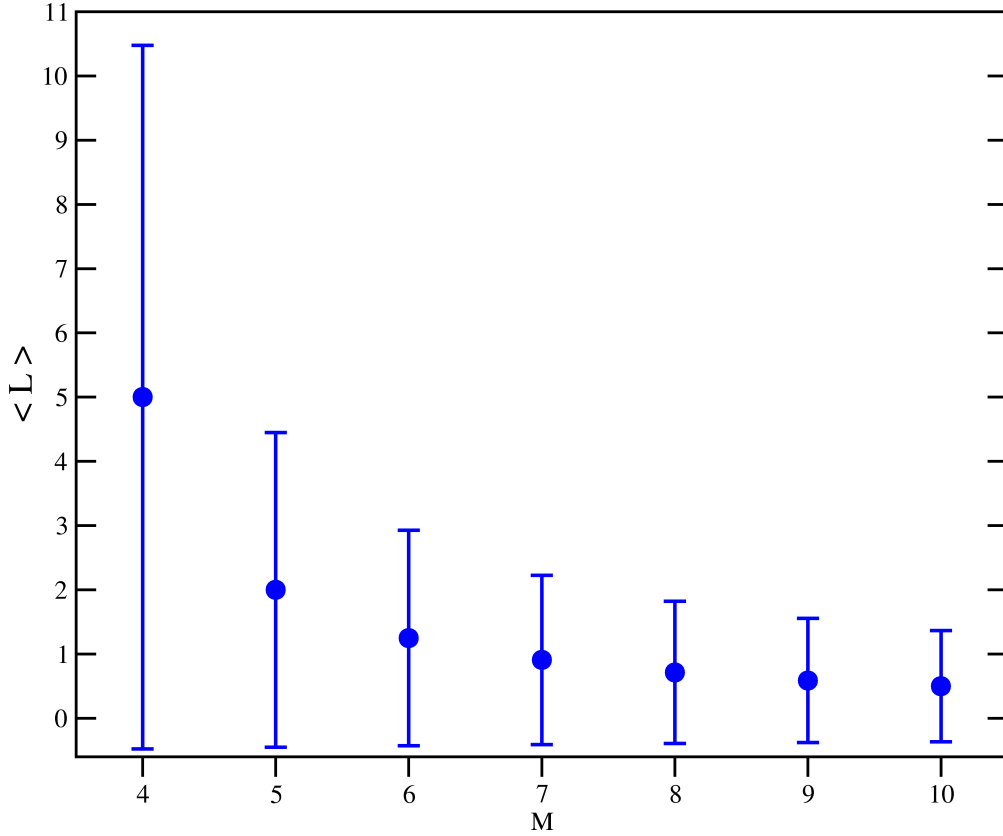


Figure 4: Queue length for  $T_S = 50$  min and  $T_C = 15$  min.

Summing up the series,

$$\sum_{k=0}^{\infty} \frac{(\alpha \rho t)^k}{k!} = e^{\alpha \rho t}, \quad (15)$$

allows recast Eq. (14) as

$$g(t) = (1 - \rho) \alpha e^{-(1-\rho) \alpha t}. \quad (16)$$

Thus, we obtain an exponential distribution with parameter  $(1 - \rho) \alpha = (1 - \rho) M \mu$ . In this manner, when all servers are busy, the mean waiting time of a patient in the row before being served is

$$\langle T \rangle = \frac{1}{(1 - \rho) M \mu} = \frac{\langle L \rangle}{\lambda}. \quad (17)$$

The result  $\langle L \rangle = \lambda \langle T \rangle$  is the expression of the well known Little's Law [7, 8]. Our derivation is an alternative statistical approach in the stationary framework [20].

We define the level of service (LOS) as the fraction of patients served in a time less than a predefined threshold of quality  $T_{\text{LOS}}$ . When a new call arrives, if there are idle servers, an ambulance is dispatched and there is not waiting time, but under full occupation, the quality threshold is fulfilled with probability

$$P(t < T_{\text{LOS}}) = \int_0^{T_{\text{LOS}}} g(t) dt = 1 - e^{-(1-\rho) \alpha T_{\text{LOS}}}. \quad (18)$$

Thus,  $\text{LOS} = (1 - P(\text{occup})) 1 + P(\text{occup}) P(t < T_{\text{LOS}}) = 1 - P(\text{occup}) (1 - P(t < T_{\text{LOS}}))$ , and in our case we obtain,

$$\text{LOS} = 1 - P(\text{occup}) e^{-(1-\rho) M \mu T_{\text{LOS}}}, \quad (19)$$

where  $P(\text{occup})$  is given by Eq. (9). Then,  $\text{LOS} = 1$  if and only if  $P(\text{occup}) = 0$  which is the desired condition for management of emergency calls. In Figure 5, we illustrate the dependence of LOS in the number of ambulances.

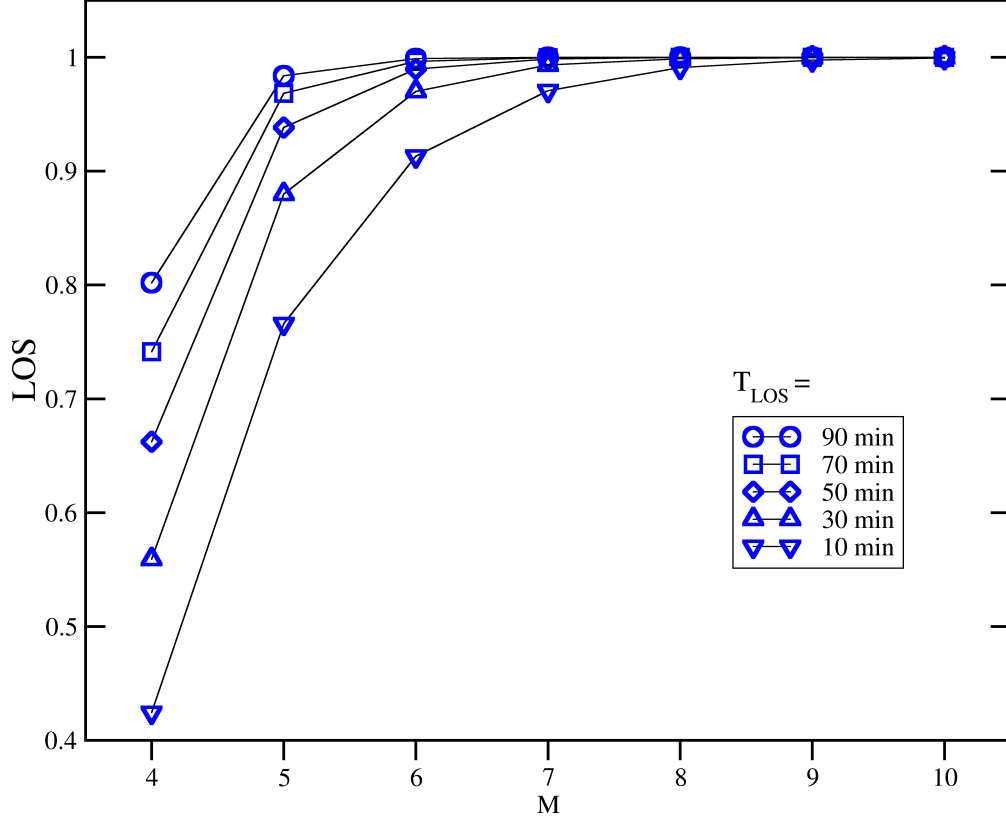


Figure 5: Level of service for  $T_S = 50$  min and  $T_C = 15$  min and different values of  $T_{\text{LOS}}$ . Lines between points are only to guide the eye.

#### 4.4 Performance of servers

The probability that a given server is busy can be written as

$$P(\text{busy}) = \sum_{n=1}^{\infty} P(\text{busy}|n) \pi_n, \quad (20)$$

where  $\pi_n$  is given by Eq. (6) and  $P(\text{busy}|n)$  is the conditional probability of a server busy given that the state of the system is  $n$ . For  $n < M$  and under the assumption that the assignment of calls to servers is at random if there is more than one idle, using simple combinatorial calculation, results

$$P(\text{busy}|n) = \frac{\binom{M-1}{n-1}}{\binom{M}{n}} = \frac{n}{M}, \quad (21)$$

whereas, for  $n \geq M$ ,  $P(\text{busy}|n) = 1$ . Then,

$$P(\text{busy}) = \sum_{n=1}^{M-1} \frac{n}{M} \pi_n + P(\text{occup}). \quad (22)$$



Therefore, from Eqs. (6) and (9) results

$$P(\text{busy}) = \frac{1}{S} \left( \sum_{n=1}^{M-1} \frac{M^{n-1}}{(n-1)!} \rho^n + \frac{M^M}{M!} \frac{\rho^M}{(1-\rho)} \right). \quad (23)$$

In the stationary state,  $P(\text{busy})$  is the fraction of time that a given server remains busy. The dependence on  $M$  of  $P(\text{busy})$  and  $P(\text{occup})$  is shown in Figure 6 for our example. In this situation,

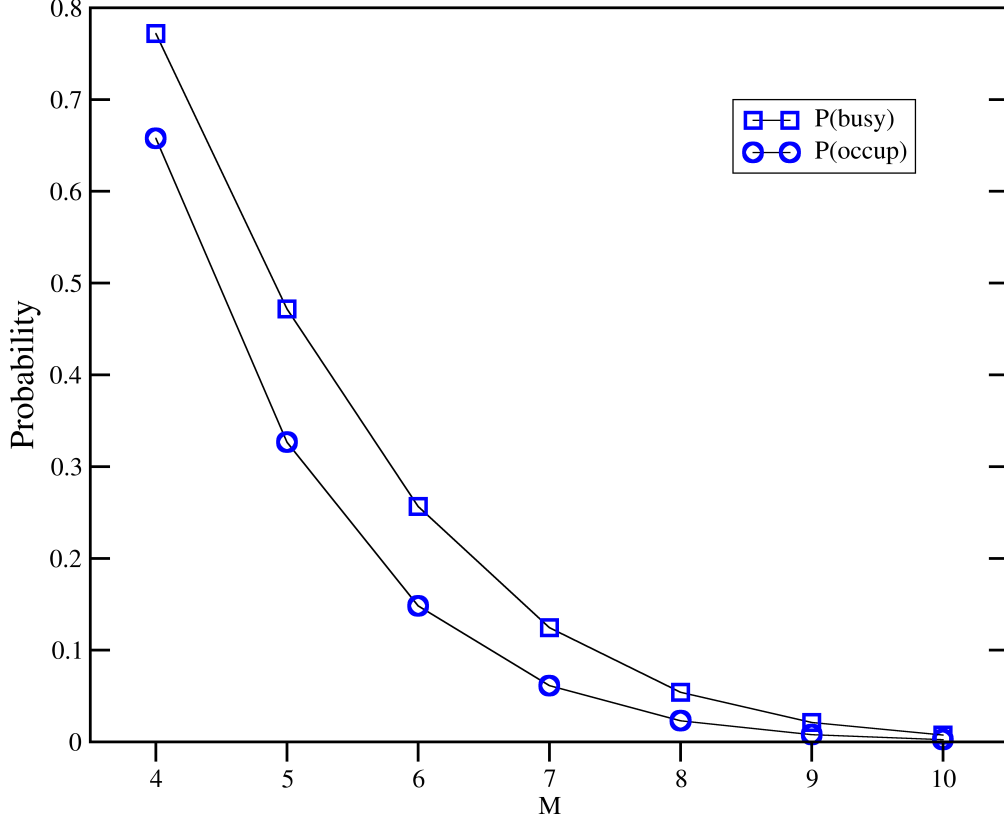


Figure 6:  $P(\text{busy})$  and  $P(\text{occup})$  for  $T_S = 50$  min and  $T_C = 15$  min. Lines between points are only to guide the eye.

from Figure 5 and 6, we can see that a fleet of six ambulances implies LOS greater than 90% even though we choose small values of  $T_{\text{LOS}}$ . However, a fleet of six ambulances remains completely occupied only 14.8% of time and each server is busy only 25.7% of time. To put in clear the trade-off between level of service and use of resources, we plot in Figure 7 the log-log plane of probability ( $P(\text{busy})$  or  $P(\text{occup})$ ) and  $(1 - \text{LOS})$  as function of  $M$ . The graph is constructed for the fixed value  $T_{\text{LOS}} = 30$  min.

Alternatively, we can look at the mean number of medical attentions given by the system per unit of time,

$$\langle a \rangle = \sum_{n=1}^{\infty} \omega_n^- \pi_n. \quad (24)$$

From Eqs. (2) and (8), we obtain

$$\langle a \rangle = \mu \left( \sum_{n=1}^{M-1} n \pi_n + M P(\text{occup}) \right) = \mu M P(\text{busy}). \quad (25)$$

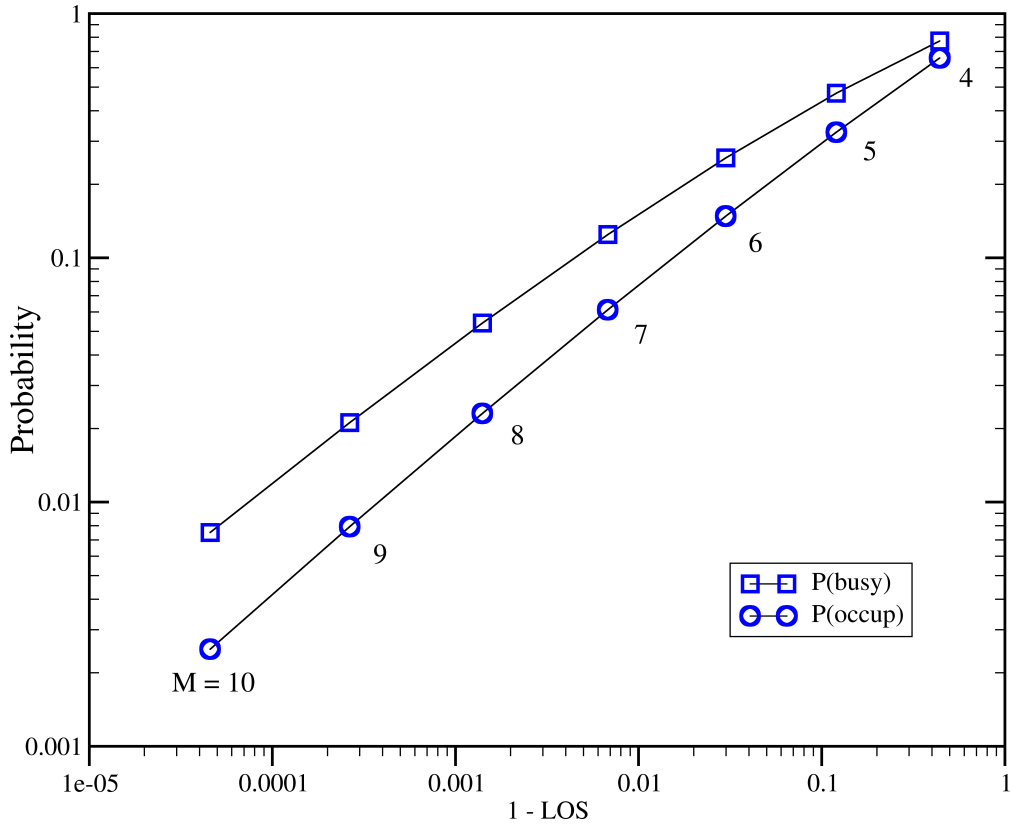


Figure 7: Log-log plane of probability ( $P(\text{busy})$  or  $P(\text{occup})$ ) and  $(1 - \text{LOS})$  as function of  $M$  for  $T_S = 50$  min,  $T_C = 15$  min and,  $T_{\text{LOS}} = 30$  min. Lines between points are only to guide the eye.

In this way, if all servers are equivalents, the simple unit hour utilization, is given by  $\mu P(\text{busy})$ . Last, but not least, if server's cost per attention is  $C$ , the cost per ambulance in the same unit of time results  $C \mu P(\text{busy})$ .

## 5 Concluding remarks

This work considers the mathematical problem of the number of ambulances needed in operation for a EMS according to the mean time between entrance calls and service times. We developed our analysis in the framework of the queueing theory for the nonstationary as well as for the stationary regime.

In Section 3 we presented a novel use of a previous result of MFPT for calculating the average time to the next critical condition of full occupation of servers. Our description allows to cope with the problem of the number of ambulances needed to avoid that condition. In Section 4 we rederived standard results of queueing theory for the stationary state emphasizing the use of conditional probabilities in the saturation regime when all servers are busy. We have paid special attention to the key performance indicators as queue length, level of service, and fraction of time of use of servers. Our analysis allows stress in simple mathematical terms the trade-off between quality of service and use of servers as function of fleet size.

## 6 Appendix: Mathematics of the random walk

### 6.1 MFPT

For asymmetric and site dependent transition probabilities, the analytical expressions for the MFPT of a random walk with a reflecting boundary condition, as shown in Figure 1, is given by [17],

$$\begin{aligned} T(0) &= \sum_{k=0}^M \frac{1}{w_k^+} + \sum_{k=0}^{M-1} \frac{1}{w_k^+} \sum_{i=k+1}^M \prod_{j=k+1}^i \frac{w_j^-}{w_j^+}, \\ T(1) &= T(0) - \frac{1}{w_0^+}, \\ T(n) &= T(0) - \sum_{k=0}^{n-1} \frac{1}{w_k^+} - \sum_{k=0}^{n-2} \frac{1}{w_k^+} \sum_{i=k+1}^{n-1} \prod_{j=k+1}^i \frac{w_j^-}{w_j^+} \quad (2 \leq n \leq M). \end{aligned} \quad (26)$$

In our model with  $M$  servers, using Eq. (2) and the parameter  $\gamma$  defined in the text, we can recast the products in Eq. (26) as

$$\prod_{j=k+1}^i \frac{w_j^-}{w_j^+} = \gamma^{i-k} \prod_{j=k+1}^i j = \gamma^{i-k} \frac{i!}{k!}. \quad (27)$$

Thus, we can also recast the sums in Eq. (26) as

$$\sum_{i=k+1}^{n-1} \prod_{j=k+1}^i \frac{w_j^-}{w_j^+} = \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i! \gamma^i, \quad (28)$$

and

$$\sum_{k=0}^{n-2} \frac{1}{w_k^+} \sum_{i=k+1}^{n-1} \prod_{j=k+1}^i \frac{w_j^-}{w_j^+} = \frac{1}{\lambda} \sum_{k=0}^{n-2} \frac{\gamma^{-k}}{k!} \sum_{i=k+1}^{n-1} i! \gamma^i. \quad (29)$$

Replacing the last two expressions in Eq. (26), we obtain Eq. (3) in Sec. 3 in the main text.

### 6.2 Steady state

Following [10], we can construct the steady state of the problem. From Eq. (1), the time independent solution must satisfy

$$\begin{aligned} \omega_1^- \pi_1 - \omega_0^+ \pi_0 &= 0, \\ \omega_{n+1}^- \pi_{n+1} + \omega_{n-1}^+ \pi_{n-1} - (\omega_n^+ + \omega_n^-) \pi_n &= 0 \quad (n \geq 1). \end{aligned} \quad (30)$$

Thus, from the first expression we immediately obtain

$$\pi_1 = \frac{\omega_0^+}{\omega_1^-} \pi_0. \quad (31)$$

Substituting this result in the second equation ( $n = 1$ ) yields

$$\pi_2 = \frac{\omega_1^+ \omega_0^+}{\omega_2^- \omega_1^-} \pi_0, \quad (32)$$

and so on we can proof by induction that

$$\pi_n = \frac{\omega_{n-1}^+ \dots \omega_0^+}{\omega_n^- \dots \omega_1^-} \pi_0 = \prod_{j=1}^n \frac{\omega_{j-1}^+}{\omega_j^-} \pi_0 \quad (n \geq 1). \quad (33)$$

From the normalization condition,  $\sum_{n=0}^{\infty} \pi_n = 1$ , results

$$\pi_0 = \frac{1}{S}, \quad (34)$$

where

$$S = 1 + \sum_{n=1}^{\infty} \prod_{j=1}^n \frac{\omega_{j-1}^+}{\omega_j^-}. \quad (35)$$

The existence of the steady state is determined by the convergence of the series in the Eq. (35).

For the model given by Eq. (2), the product in Eq. (35) can be written as,

$$\prod_{j=1}^n \frac{\omega_{j-1}^+}{\omega_j^-} = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n \prod_{j=1}^n \frac{1}{j} = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} & (0 \leq n \leq M), \\ \left(\frac{\lambda}{\mu}\right)^n \prod_{j=1}^{M-1} \frac{1}{j} \prod_{j=M}^n \frac{1}{M} = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{(M-1)!} \frac{1}{M^{n-(M-1)}} & (n \geq M). \end{cases} \quad (36)$$

Substituting Ec. (36) into Eq. (35) we obtain,

$$\begin{aligned} S &= 1 + \sum_{n=1}^{M-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \sum_{n=M}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{(M-1)!} \frac{1}{M^{n-(M-1)}}, \\ &= \sum_{n=0}^{M-1} \frac{(M\rho)^n}{n!} + \frac{M^{M-1}}{(M-1)!} \sum_{n=M}^{\infty} \rho^n, \end{aligned} \quad (37)$$

where  $\rho$  is given by Eq. (5). The convergence of series in the last expression only occurs for  $\rho < 1$ . In this case,

$$\sum_{n=M}^{\infty} \rho^n = \sum_{i=0}^{\infty} \rho^{i+M} = \rho^M \sum_{i=0}^{\infty} \rho^i = \frac{\rho^M}{1-\rho}, \quad (38)$$

and from Eqs. (33), (36), and (37), (38), we obtain the Eqs. (6) and (7) in the main text, respectively.

## Acknowledgment

An early stage of this work was partially supported by *Sistema de Urgencias del Rosafe SA*, Córdoba, Argentina.

## References

- [1] Jeffrey B. Goldberg. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, 1(1):20–39, 2004.
- [2] C. Lakshmi and Appa Iyer Sivakumar. Application of queueing theory in health care: A literature review. *Operations Research for Health Care*, 2(1-2):25–39, 2013.
- [3] William K. Hall. Management science approaches to the determination of urban ambulance requirements. *Socio-Econ. Plan. Sci.*, 5(5):491–499, 1971.
- [4] Geoffrey N. Berlin and Jon C. Liebman. Mathematical analysis of emergency ambulance location. *Socio-Econ. Plan. Sci.*, 8(6):323–328, 1974.

- [5] Renata Algisi Takeda, João A. Widmer, and Reinaldo Morabito. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*, 34(3):727–741, 2007.
- [6] Marcos Singer and Patricio Donoso. Assessing an ambulance service with queueing theory. *Computers & Operations Research*, 35(8):2549–2560, 2008.
- [7] John D. C. Little. A proof for the queueing formula:  $l = \lambda w$ . *Operations Research*, 9(3):383–387, 1961.
- [8] John D. C. Little. Little’s law as viewed on its 50th anniversary. *Operations Research*, 59(3):536–549, 2011.
- [9] Ger Koole and Avishai Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1-4):41–59, 2002.
- [10] Wayne L. Winston. *Operations Research: Applications and Algorithms*. Duxbury Press, 4 ed. edition, 2003. Chap. 20.
- [11] David R. Cox and Walter L. Smith. *Queues*. Methuen, London, 1961.
- [12] Sidney Redner. *A Guide to First-Passage Processes*. Cambridge University Press, Cambridge, U.K., 2001.
- [13] David S. Matteson, Mathew W. McLean, Dawn B. Woodard, and Shane G. Henderson. Forecasting emergency medical service call arrival rates. *Ann. Appl. Stat.*, 5(2B):1379–1406, 2011.
- [14] M.S. Eisenberg, L. Bergner, and A. Hallstrom. Cardiac resuscitation in the community: Importance of rapid provision and implications for program planning. *JAMA: The Journal of the American Medical Association*, 241(18):1905–1907, 1979.
- [15] Stan Feero, Jerris R. Hedges, Erik Simmons, and Lisa Irwin. Does out-of-hospital EMS time affect trauma survival? *The American Journal of Emergency Medicine*, 13(2):133–135, 1995.
- [16] Peter T Pons and Vincent J Markovchick. Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine*, 23(1):43–48, 2002.
- [17] Pedro A. Pury and Manuel O. Cáceres. Mean first-passage and residence times of random walks on asymmetric disordered chains. *Journal of Physics A: Mathematical and General*, 36(11):2695–2706, 2003.
- [18] Hubert Setzler, Cem Saydam, and Sungjune Park. EMS call volume predictions: A comparative study. *Computers & Operations Research*, 36(6):1843–1851, 2009.
- [19] Sheldon M. Ross. *Simulation*. Elsevier Academic Press, 4 ed. edition, 2006. Pag. 31.
- [20] Song-Hee Kim and Ward Whitt. Statistical analysis with little’s law. *Operations Research*, 61(4):1030–1045, 2013.